



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: Masataka Andoh, et al.

Examiner: Unassigned

Serial No.: 10/697,762

Group Art Unit: Unassigned

Filed: October 30, 2003

Docket: 17199

For: EDR DIRECTION ESTIMATING
METHOD, SYSTEM, AND PROGRAM
AND MEMORY MEDIUM FOR STORING
THE PROGRAM

Dated: December 1, 2003


**Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450**

CLAIM OF PRIORITY

Sir:

Applicants in the above-identified application hereby claim the right of priority in connection with Title 35 U.S.C. §119 and in support thereof, herewith submit a certified copy of Japanese Patent Application 2003-049223, filed on February 26, 2003.

Respectfully submitted,


Paul J. Esatto, Jr.
Registration No. 30,749

Scully, Scott, Murphy & Presser
400 Garden City Plaza
Garden City, New York 11530
(516) 742-4343
PJE:ahs

CERTIFICATE OF MAILING UNDER 37 C.F.R. §1.8(a)

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner For Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on December 1, 2003.

Dated: December 1, 2003


Paul J. Esatto, Jr.

US

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 2 月 2 6 日
Date of Application:

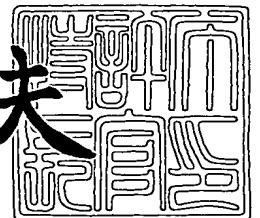
出 願 番 号 特 願 2 0 0 3 - 0 4 9 2 2 3
Application Number:
[ST. 10/C]: [J P 2 0 0 3 - 0 4 9 2 2 3]

出 願 人 日 本 電 気 株 式 有 限 公 司
Applicant(s): 大 瀧 慈

2 0 0 3 年 9 月 8 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 7 3 3 3 4

【書類名】 特許願

【整理番号】 64002113

【特記事項】 特許法第 3 0 条第 1 項の規定の適用を受けようとする特
許出願

【提出日】 平成15年 2月26日

【あて先】 特許庁長官殿

【国際特許分類】 G06N 7/00

【発明者】

 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

 【氏名】 安東 正貴

【発明者】

 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

 【氏名】 斎藤 彰

【発明者】

 【住所又は居所】 広島県廿日市市宮園 9 丁目 1 の 7

 【氏名】 大瀧 慈

【発明者】

 【住所又は居所】 広島県広島市佐伯区楽々園 5 - 9 五日市住宅 1 6 - 2
0 2

 【氏名】 佐藤 健一

【特許出願人】

 【識別番号】 000004237

 【氏名又は名称】 日本電気株式会社

【特許出願人】

 【住所又は居所】 広島県廿日市市宮園 9 丁目 1 の 7

 【氏名又は名称】 大瀧 慈

【代理人】**【識別番号】** 100071272**【弁理士】****【氏名又は名称】** 後藤 洋介**【選任した代理人】****【識別番号】** 100077838**【弁理士】****【氏名又は名称】** 池田 憲保**【手数料の表示】****【予納台帳番号】** 012416**【納付金額】** 21,000円**【提出物件の目録】****【物件名】** 明細書 1**【物件名】** 図面 1**【物件名】** 要約書 1**【包括委任状番号】** 0018587**【プルーフの要否】** 要

【書類名】 明細書

【発明の名称】 E D R 方向推定方法、システム、プログラム、及び記録媒体

【特許請求の範囲】

【請求項 1】 大量変数に関する単一指標モデルにおいて、E D R 方向を推定する E D R 方向推定方法において、

解析対象となるデータファイルを入力するステップと、

目的変数と説明変数の組からなる解析対象データを受け、前記説明変数を基準化して、基準化された説明変数と前記目的変数の組からなるデータを出力するステップと、

前記基準化された説明変数と前記目的変数の組からなるデータを受け、該データを前記目的変数の所定の閾値を基準として 2 つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するステップと、

該各平均ベクトルを受け、当該 2 つの平均ベクトルの差を計算して E D R 方向を求め、該 E D R 方向データを前記データ変換手段へ出力するステップと、

当該 E D R 方向データを単位ベクトルに変換し、その単位ベクトルを E D R 方向推定値として出力するステップ

を有することを特徴とする E D R 方向推定方法。

【請求項 2】 前記 E D R 方向を計算するステップにおいて、

相関行列の逆行列が存在する場合に、相関行列の逆行列で前記 E D R 方向データを補正し、前記 E D R 方向データ及び前記補正した E D R 方向データを前記データ変換手段へ送出し、

前記相関行列の逆行列が存在しない場合、前記 E D R 方向データのみ前記データ変換手段へ送出することを特徴とする請求項 1 記載の E D R 方向推定方法。

【請求項 3】 前記閾値を前記目的変数の中央値とすることを特徴とする請求項 1 又は 2 記載の E D R 方向推定方法。

【請求項 4】 前記閾値を前記目的変数の平均値とすることを特徴とする請求項 1 又は 2 記載の E D R 方向推定方法。

【請求項 5】 前記目的変数が 2 値である場合、前記閾値を 0. 5 とするこ

とを特徴とする請求項 1 又は 2 記載の E D R 方向推定方法。

【請求項 6】 前記説明変数の基準化の際、前記基準化された説明変数をスライスに分割する際、前記平均ベクトルを計算する際に、欠測値を取り除いて計算することを特徴とする請求項 1 ～ 5 のいずれかに記載の E D R 方向推定方法。

【請求項 7】 解析対象となるデータファイルを入力する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含み、大量変数に関する単一指標モデルにおいて、E D R 方向を推定する E D R 方向推定システムにおいて、

前記データ解析装置は、

前記入力装置から、目的変数と説明変数の組からなる解析対象データを受け取り、前記説明変数を基準化して、基準化された説明変数と前記目的変数の組からなるデータを出力するデータ変換手段と、

前記基準化された説明変数と前記目的変数の組からなるデータを入力し、該データを前記目的変数の所定の閾値を基準として 2 つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するスライス平均計算手段と、

該各平均ベクトルを入力し、当該 2 つの平均ベクトルの差を計算して E D R 方向を求め、該 E D R 方向データを前記データ変換手段へ出力する E D R 方向計算手段とを有し、

前記データ変換手段は、当該 E D R 方向データを単位ベクトルに変換し、その単位ベクトルを E D R 方向推定値として前記出力装置に出力する

ことを特徴とする E D R 方向推定システム。

【請求項 8】 前記 E D R 方向計算手段は、相関行列の逆行列が存在する場合に、相関行列の逆行列で前記 E D R 方向を補正し、前記 E D R 方向データ及び前記補正した E D R 方向データを前記データ変換手段へ送出し、前記相関行列の逆行列が存在しない場合、前記 E D R 方向データのみ前記データ変換手段へ送出することを特徴とする請求項 7 記載の E D R 方向推定システム。

【請求項 9】 大量変数に関する単一指標モデルにおいて、E D R 方向を推定するためコンピュータに、

解析対象となるデータファイルを入力するステップと、

目的変数と説明変数の組からなる解析対象データを受け、前記説明変数を基準化して、当該基準化された説明変数と前記目的変数の組からなるデータを出力するステップと、

前記基準化された説明変数と前記目的変数の組からなるデータを受け、該データを前記目的変数の所定の閾値を基準として2つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するステップと、

該各平均ベクトルを受け、当該2つの平均ベクトルの差を計算してEDR方向を求め、該EDR方向データを前記データ変換手段へ出力するステップと、

当該EDR方向データを単位ベクトルに変換し、その単位ベクトルをEDR方向推定値として出力するステップ

を実行させるためのEDR方向推定プログラム。

【請求項10】 大量変数に関する単一指標モデルにおいて、EDR方向を推定するためコンピュータに、

解析対象となるデータファイルを入力するステップと、

目的変数と説明変数の組からなる解析対象データを受け、前記説明変数を基準化して、当該基準化された説明変数と前記目的変数の組からなるデータを出力するステップと、

前記基準化された説明変数と前記目的変数の組からなるデータを受け、該データを前記目的変数の所定の閾値を基準として2つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するステップと、

該各平均ベクトルを受け、当該2つの平均ベクトルの差を計算してEDR方向を求め、該EDR方向データを前記データ変換手段へ出力するステップと、

当該EDR方向データを単位ベクトルに変換し、その単位ベクトルをEDR方向推定値として出力するステップ

を実行させるためのEDR方向推定プログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】**【0 0 0 1】****【発明の属する技術分野】**

本発明は単一指標モデルにおける E D R 方向の推定方法およびシステムに関し、特に大量変数に関する単一指標モデルにおける E D R 方向の推定方法、システム、プログラム、及び記録媒体に関する。

【0 0 0 2】**【従来の技術】**

一般に、実際の現象を統計的に解析する目的の 1 つは、種々の特性間の関係を見出し、予測を行うことである。このような場合、回帰分析を用いてデータから何らかの関係を見出し、ある変数に対して予測することがよく行われる。例えば、線形回帰分析やロジステック回帰分析などを用いて、目的変数 y と説明変数 x の関係を解析する。

しかし、説明変数 x の次元 p が大きくなればなるほど、この種の回帰分析をすることが困難になる。この問題を解決するために、説明変数の次元数を減少させる方法がいくつか考案されている。

例えば、以下の非特許文献 1 を参照すると、K e r - C h a u L i は S I R (Sliced Inverse Regression) を考案した。

【0 0 0 3】

S I R は説明変数の次元数を減少するために、目的変数 y を説明するのに十分な x の部分空間を求める方法である。ここで、求めた部分空間を E D R 空間と呼び、E D R (Effective Dimension Reduction) 空間を張るベクトルのことを E D R 方向ベクトルと呼ぶ。この次元数が減少した E D R 空間において通常の回帰分析を行うことにより、目的変数 y と説明変数 x の関係を調べることができる。

【0 0 0 4】

又、以下の非特許文献 2 を参照すると、H a l l と I c h i m u r a は平滑化法を用いて E D R 方向を推定した。

【0 0 0 5】

又、以下の非特許文献 3 を参照すると、X i a 等是非線形平滑化法を用いた E

DR空間を推定する手法を提案した。しかし、説明変数の数が膨大になると計算が非常に困難となる。

【0006】

次に、SIRについて説明する。SIRにおいては、以下の数1～数6に示されるようなようなモデルを仮定している。

【0007】

【数1】

$$y = f(\beta_1'x, \dots, \beta_k'x, \varepsilon)$$

ここで、変数 y は目的変数とし、 f は未知関数とし、 ε は x と独立な確率変数とし、 x は p 次元の説明変数とする。また、 β_1, \dots, β_k は p 次元の未知係数ベクトルとし、EDR方向ベクトルとする。

【0008】

図8、図9を用いて、SIRを説明する。初めに、入力装置1により入力されたデータファイルの説明変数をデータ解析装置2のデータ基準化手段24により基準化する（図9のステップA1）。

【0009】

【数2】

$$z_i = \sum_{xx}^{-\frac{1}{2}} [x_i - \bar{x}] \quad (i = 1, \dots, n)$$

ただし、 z_i はそれぞれ、 x の分散共分散行列、平均である。

【0010】

【外1】

$$\sum_{xx} \bar{x}$$

次に、スライス平均計算手段22により、目的変数 y をソートし、 H 個のスライス I_1, \dots, I_H に分割する（ステップA2）。スライス I_k に属する目

的変数の割合を外 1 - 1 として計算する（以下の数 3 参照）。

【 0 0 1 1 】

【外 1 - 1】

\hat{p}_k

【数 3】

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^n \delta_k(y_i)$$

ここで、 $\delta_k(y_i)$ は外 2 とする。

【 0 0 1 2 】

【外 2】

$$\delta_k(y_i) = \begin{cases} 1, & y_i \in I_k, \\ 0, & y_i \notin I_k. \end{cases}$$

次に、以下の数 4 に示される数式を用いて、スライスごとに、基準化された説明変数の平均ベクトルを計算する（ステップ A 3）。

【 0 0 1 3 】

【数 4】

$$m_k = \left[\frac{1}{n\hat{p}_k} \right] \sum_{y_i \in I_k} z_i$$

次に、主成分分析手段 2 5 により、スライスごとの平均ベクトル m に対して主成分分析を行い、固有ベクトルを求める（ステップ A 4）。

【 0 0 1 4 】

ここで、以下の数 5 に示される数式を用いて固有値・固有ベクトルを求める。

【 0 0 1 5 】

【数 5】

$$V = \sum_{k=1}^H \hat{p}_k m_k m'_k$$

データ基準化手段 24 により、個有値が大きい方から K 個の固有ベクトル η_k ($k = 1, \dots, K$) を抽出し、以下の数 6 に示された数式を用いて、元の座標系に変換する (ステップ A5)。

【0016】

【数 6】

$$\beta_k = \sum_{xx} -\frac{1}{2} \eta_k$$

出力装置 3 において、ステップ A5 で求めた EDR 方向ベクトルを出力する (ステップ A6)。

【0017】

【非特許文献 1】

K e r - C h a u L i, 1991 年、ジャーナル・オブ・ジ・アメリカン・スタティスティカル・アソシエーション、第 86 巻、第 414 号 (Journal of the American Statistical Association, vol86, 316-342, 1991)

【0018】

【非特許文献 2】

I c h i m u r a 他, 1993 年、ジ・アナルズ・オブ・スタティスティクス、第 21 巻 (The Annals of Statistics, 21, 157-178)

【0019】

【非特許文献 3】

X i a e t. a l, 2002 年、ジャーナル・オブ・ジ・ロイアル・スタティスティカル・ソサイエティ・シリーズ・ビー、第 64 巻 (Journal of the Royal Statistical Society : Series B)

【0020】

【発明が解決しようとする課題】

上記した従来技術における第1の問題点は、SIRが遺伝子発現解析用DNAチップやマイクロアレイなどの大量の変数を持つデータに適用できないことである。SIRにおいてはデータを基準化するために、説明変数の分散共分散行列の逆行列を必要としたり、EDR方向ベクトルを推定するために主成分分析を行って、固有ベクトルを求めたりする必要がある。しかし、大量変数においては、分散共分散行列の逆行列が数値計算上求められなかったり、主成分分析における計算時間が膨大となったりする。

第2の問題点は、SIRは説明変数の分布を楕円分布に限定していることである。そのため、説明変数が2値の場合には適用することができなかった。

【0021】

本発明の目的は、次の式で表される単一指標モデル (Single Index Model) に関して、スライス数が2つのときに分散共分散行列の逆行列および主成分分析を用いずに、単純な計算でEDR方向を推定する方法およびシステムを提供することである。単一指標モデルとは、1つの未知係数ベクトルからなるモデルであり、従来の重回帰分析やロジスティック回帰分析などを包含するようなモデルである。

【0022】

ここで、単一指標モデルは以下の数7に示されるような数式で表すことができる。

【0023】**【数7】**

$$y = f(\beta_0' x, \varepsilon)$$

ここで、変数 y は目的変数とし、 f は未知の大局的な単調関数とし、 ε は x と独立な確率変数とし、 x は p 次元の説明変数とする。また、 β_0 は p 次元の未知係数ベクトルとし、真のEDR方向ベクトルとする。

本発明の他の目的は、説明変数 x に特定の分布を仮定しないことである。これ

により、本発明のEDR方向の推定システムが、説明変数が2値の場合にも適用できるようにすることにある。

本発明の更に他の目的は、大量変数のデータである、遺伝子発現解析用DNAチップやマイクロアレイなどのデータから、重要な遺伝子を探索する手法およびシステムを提供することである。

【0024】

【課題を解決するための手段】

本発明に係るEDR方向推定システムは、
解析対象となるデータファイルを入力する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含み、

前記データ解析装置は、

前記入力装置から、目的変数と説明変数の組からなる解析対象データを受け取り、前記説明変数を基準化して、基準化された説明変数と前記目的変数の組からなるデータを出力するデータ変換手段と、

前記基準化された説明変数と前記目的変数の組からなるデータを入力し、該データを前記目的変数の所定の閾値を基準として2つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するスライス平均計算手段と、

該各平均ベクトルを入力し、当該2つの平均ベクトルの差を計算してEDR方向を求め、該EDR方向データを前記データ変換手段へ出力するEDR方向計算手段とを有し、

前記データ変換手段は、当該EDR方向データを単位ベクトルに変換し、その単位ベクトルをEDR方向推定値として前記出力装置に出力する

ことを特徴とする。

【0025】

又、本発明に係るEDR方向推定方法は、
解析対象となるデータファイルを入力するステップと、
目的変数と説明変数の組からなる解析対象データを受け、前記説明変数を基準化して、基準化された説明変数と前記目的変数の組からなるデータを出力するス

テップと、

前記基準化された説明変数と前記目的変数の組からなるデータを受け、該データを前記目的変数の所定の閾値を基準として2つのスライスに分割して、各スライスごとに前記基準化された説明変数の平均ベクトルを計算し、該各平均ベクトルを出力するステップと、

該各平均ベクトルを受け、当該2つの平均ベクトルの差を計算してEDR方向を求め、該EDR方向データを前記データ変換手段へ出力するステップと、

当該EDR方向データを単位ベクトルに変換し、その単位ベクトルをEDR方向推定値として出力するステップ

を有することを特徴とする

【0026】

【発明の実施の形態】

次に、本発明の第1の実施の形態について図面を参照して詳細に説明する。図1を参照すると、本発明の第1の実施の形態は、解析対象となるデータファイルを入力する入力装置1と、プログラム制御により動作するデータ解析装置2と、ディスプレイ装置や印刷装置等の出力装置3とを含む。解析対象となるデータファイルはN個のデータの組からなり、それぞれの組は、1つの目的変数とp次元の説明変数からなる。データ解析装置2は、データ変換手段21と、スライス平均計算手段22と、EDR方向計算手段23とを備えている。

【0027】

データ変換手段21は、与えられたデータファイルのN個のp次元説明変数を基準化して、基準化された説明変数と目的変数の組からなるデータをスライス平均計算手段22に送る。また、EDR方向計算手段23から与えられたEDR方向と、補正したEDR方向とを元の座標系に変換し、さらに単位ベクトルに変換して出力装置3へ送る。

【0028】

スライス平均計算手段22は、目的変数の中央値を基準にして、N個のデータの組を2つのスライスに分割する。さらに、それぞれのスライスにおいて、基準化されたp次元の説明変数の平均ベクトルを計算して、EDR方向計算手段23

へ送る。

【0029】

E D R 方向計算手段 23 は、スライス平均計算手段 22 から与えられた 2 つの平均ベクトルの差を求める。これが E D R 方向である。また、p 次元説明変数の相関行列を求め、相関行列の逆行列が計算できれば、相関行列の逆行列で E D R 方向を補正し、E D R 方向と補正した E D R 方向をデータ変換手段 21 へ送る。一方、相関行列の逆行列が計算できなければ、E D R 方向のみをデータ変換手段 21 へ送る。

【0030】

次に、図 1 及び図 2 を参照して本実施の形態の動作について詳細に説明する。解析対象となるデータファイルにおけるデータは以下の数 8 に示されたものとする。

【0031】

【数 8】

$$(y_i, x_i), \quad i = 1, \dots, N$$

ここで、 y_i を目的変数とし、 x_i を p 次元の説明変数とする。解析対象データはデータ変換手段 21 へ送られる。データ変換手段 21 は、説明変数のサンプル平均外 3 及び分散外 4 により、以下の数 9 に示されるように説明変数 x_i (j) を基準化する。

【0032】

【外 3】

$$\bar{\mu}^{(j)}$$

【外 4】

$$(\bar{\sigma}^{(j)})^2$$

【数 9】

$$z_i^{(j)} = \frac{x_i^{(j)} - \bar{\mu}^{(j)}}{\hat{\sigma}^{(j)}}$$

ここで、 $x_i = (x_i^{(1)}, \dots, x_i^{(p)})'$ とし、サンプル平均外 5 及び分散外 6 は、それぞれ以下の数 1 0 及び数 1 1 とする（図 2 のステップ A 1）。

【0 0 3 3】

【外 5】

 $\bar{\mu}^{(j)}$

【外 6】

 $(\hat{\sigma}^{(j)})^2$

【数 1 0】

$$\bar{\mu}^{(j)} = \frac{\sum_{i=1}^N x_i^{(j)}}{N}$$

【数 1 1】

$$(\hat{\sigma}^{(j)})^2 = \frac{\sum_{i=1}^N (x_i^{(j)} - \bar{\mu}^{(j)})^2}{N - 1}$$

スライス平均計算手段 2 2 は、解析対象データにおける目的変数 y_i を以下の数 1 2 に示される数式に従って、2 つのスライス I_H と I_L に分割する。

【0 0 3 4】

【数 1 2】

$$I_H = \{i | y_i \geq t, i \in I\}, \quad I_L = \{i | y_i < t, i \in I\}$$

ここで、閾値 t は y の中央値とし、 $I = \{1, \dots, N\}$ とする（ステップ A 2）。次に、それぞれのスライス I_H , I_L に対して、基準化された説明変数 z_i の平均ベクトル外 7, 外 8 を以下の数 1 3 に示された数式に従って計算する。

【0 0 3 5】

【外 7】

$$\hat{m}_H$$

【外 8】

$$\hat{m}_L$$

【数 1 3】

$$\hat{m}_H = \frac{1}{N_H} \sum_{i \in I_H} z_i, \quad \hat{m}_L = \frac{1}{N_L} \sum_{i \in I_L} z_i$$

ここで、 N_H は I_H に属するデータの個数であり、 $N_L = N - N_H$ 、 $Z_i = (Z_i(1), \dots, Z_i(1))'$ とする（ステップ A 3）。

【0 0 3 6】

EDR 方向計算手段 2 3 は、以下の数 1 4 に示された数式に従ってステップ A 3 で求めた平均ベクトルの差を計算する（ステップ A 4）。

【0 0 3 7】

【数 1 4】

$$\hat{\eta} = \frac{1}{2} (\hat{m}_H - \hat{m}_L)$$

次に、ステップ A 5 で説明変数の相関行列外 9 を計算する。

【0 0 3 8】

【外 9】

$\hat{\Omega}$

ステップ A 6 で、相関行列外 1 0 の逆行列外 1 1 を求めることができれば、逆行列を用いて、以下の数 1 5 に示された数式に従って外 1 2 を補正する（ステップ A 7）。

【0 0 3 9】

【外 1 0】

$\hat{\Omega}$

【外 1 1】

$\hat{\Omega}^{-1}$

【外 1 2】

$\hat{\eta}$

【数 1 5】

$$\hat{\eta}_N = \hat{\Omega}^{-1} \hat{\eta}$$

一方、逆行列外 1 3 が求められなければ、ステップ A 8 へ進む。データ変換手段 2 1 は、求められた外 1 4，外 1 5 を元の座標系に変換し、以下の数 1 6 に示された数式に従って単位ベクトルに規格化する（ステップ A 8）。

【0 0 4 0】

【外 1 3】

$\hat{\Omega}^{-1}$

【外 1 4】

 $\hat{\eta}$

【外 1 5】

 $\hat{\eta}_N$

【数 1 6】

$$\left[\begin{array}{c} \hat{\Sigma}^{-\frac{1}{2}} \hat{\eta} \\ \hat{\Sigma}^{-\frac{1}{2}} \hat{\eta} \end{array} \right], \left[\begin{array}{c} \hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}_N \\ \hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}_N \end{array} \right]$$

ここで、外 1 6、外 1 7 とする。

【0 0 4 1】

【外 1 6】

$$\hat{\Sigma} = \text{diag} \left\{ \left(\hat{\sigma}^{(1)} \right)^2, \dots, \left(\hat{\sigma}^{(K)} \right)^2 \right\}$$

【外 1 7】

$$\hat{\Sigma}^{-1/2} = \text{diag} \left\{ \frac{1}{\hat{\sigma}^{(1)}}, \dots, \frac{1}{\hat{\sigma}^{(K)}} \right\}$$

求められたベクトルを EDR 方向の推定値として出力装置 3 で出力する。

【0 0 4 2】

出力装置 3 は、説明変数 x の EDR 方向外 1 8、外 1 9 への写像（スコア）外 2 0、外 2 1 に対する目的変数 y のプロットをグラフで表示したり、印刷したりする。

【0 0 4 3】

【外 18】

 $\hat{\eta}$

【外 19】

 $\hat{\eta}_N$

【外 20】

 $\hat{\eta}'x$

【外 21】

 $\hat{\eta}'_Nx$

次に、本実施の形態の効果について説明する。本実施の形態では、主成分分析をせずに EDR 方向を推定することができるので、複雑な行列計算の必要がなく、計算時間を大幅に短縮できる。また、ベクトルの平均と差を計算するだけでよいので、SIR では不可能であった大量変数のデータに対しても EDR 方向を推定することができる。

【0044】

次に、本発明の第2の実施の形態について説明する。本発明の第2の実施の形態は、スライスを分割する閾値 t を平均値とする点である。第2の実施の形態の構成は第1の実施の形態の構成と同じであるが、第1の実施の形態の動作においては、スライスを分割する（図2のステップA2）ときに、閾値 t を中央値としたが、第2の実施の形態の動作においては、閾値 t を平均値とする点が第1の実施の形態と異なっている。

【0045】

次に、本実施の形態の効果について説明する。目的変数 y の分布が大きい値と小さい値に偏っているときに、第1の実施の形態では中央値でスライスを分割することにより、両方の分布を分割できないことが考えられるが、本実施の形態で

は平均値でスライスを分割することにより、2つの偏った分布を分割することができる。

【0046】

次に、本発明の第3の実施の形態について説明する。本発明の第3の実施の形態は、目的変数が0、1の2値の場合に、スライスを分割する閾値 t を0.5とする点である。第3の実施の形態の構成は第1の実施の形態の構成と同じであるが、第1の実施の形態の動作においては、スライスを分割する（図2のステップA2）ときに、閾値 t を中央値としたが、第3の実施の形態の動作においては、閾値 t を0.5とする点が第1の実施の形態と異なっている。

【0047】

次に、本実施の形態の効果について説明する。目的変数 y が0、1の2値のときに、第1の実施の形態では中央値でスライスを分割するために、0または1でスライスを分割してしまうが、本実施の形態では0.5でスライスを分割することにより、目的変数を0と1のスライスに分割することができる。

次に、本発明の第4の実施の形態について説明する。本発明の第4の実施の形態は、欠測値に対する扱いである。第4の実施の形態の構成は第1の実施の形態の構成と同じであるが、第1の実施の形態の動作において、データを基準化したり（図2のステップA1）、スライスに分割したり（ステップA2）、各スライス内で平均ベクトルを計算したりする（ステップA3）ときに、欠測値を取り除いて計算する点が第1の実施の形態と異なっている。

【0048】

次に、本実施の形態の効果について説明する。解析対象データから欠測値の部分のみを取り除くことにより、欠測値を含んだ個体データを解析対象から取り除くことなく有効に利用し解析できる。

次に、本発明の第5の実施の形態について図面を参照して詳細に説明する。図3を参照すると、本発明の第5の実施の形態は、本発明の第1及び第2及び第3及び第4の実施の形態と同様に、入力装置、データ解析装置、出力装置を備え、更に、データ解析プログラムを記録した記録媒体4を備える。この記録媒体4は可搬形あるいは固定型のいずれであってもよく、磁気ディスク、半導体メモリ、

CD-ROMその他の記録媒体であってもよい。

【0049】

また、本手法を実行できるコンピュータプログラムを、ネットワークに接続されたコンピュータの記録装置に格納しておき、ネットワークを介して他のコンピュータに転送することもできる。本アルゴリズムを実行するコンピュータプログラムを提供する提供媒体としては、様々な形式のコンピュータに読み出し可能な媒体として頒布可能であって、特定のタイプの媒体に限定されるものではない。

データ解析プログラムは記録媒体4からデータ解析装置5に読み込まれ、データ解析装置5の動作を制御し、入力装置1から入力されたデータファイルに対して第1及び第2及び第3及び第4の実施の形態におけるデータ解析装置2による処理と同一の処理を実行する。

【0050】

【実施例】

次に、本発明の実施例を、シミュレーションの結果を参照して具体的に説明する。かかる実施例は本発明の第1の実施の形態に対応するものである。本実施例において用いたシミュレーションモデルは以下の数17に示された数式で表される。

【0051】

【数17】

$$y = \frac{1}{1 + \exp(-5\eta'_0 z)} + \varepsilon$$

ただし、 $\varepsilon \sim N(0, 0.05^2)$ とし、 η_0 、 z は以下の数18に示された数式で表され、 $\Omega(\rho)$ は以下の数19に示される数式に従って求められる。

【0052】

【数18】

$$\eta_0 = \frac{1}{\sqrt{5}} (1, \dots, 1, 0)', \quad z = (z^{(1)}, \dots, z^{(6)})' \sim N\{0, \Omega(\rho)\}$$

【数19】

$$\Omega(\rho) = \begin{pmatrix} 1 & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\rho & 0 & 0 \\ 0 & 0 & -\rho & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

ここで、 η_0 は真のEDR方向とし、 $N(0, 1)$ は平均0分散1の正規分布を表すものとする。

【0053】

図4は、このモデルによって生成されたデータ（解析対象データ）を示す散布図である。図4において、 $N=50$ 、 $\rho=0.8$ であり、 $\eta_0'z$ （横軸）に対する目的変数 y をプロットしている。すなわち、横軸に真のEDR方向 $\eta_0'z$ が、縦軸に目的変数 y がプロットされている。ここで、 $\eta_0'z$ を真のEDR方向へのスコアと呼ぶ。このデータに対して、本発明を適用する。

【0054】

図5は、目的変数を2つのスライスに分割して（図2のステップA2）、各スライス内で平均ベクトルを計算した（ステップA3）後の $z(1)$ と $z(2)$ の散布図である。○は平均ベクトル外22、外23を示し、HとLはそれぞれ対応する目的変数が中央値よりも高いか低いを表している。図5においては、6次元の説明変数 z のうち $z(1)$ と $z(2)$ だけを示している。

【0055】

【外22】

 \hat{m}_H

【外23】

 \hat{m}_L

図6は、平均ベクトルの差（ステップA4）によって推定したEDR方向外2

4 へのスコア外 2 5（横軸）に対する目的変数 y の散布図である。尚、横軸に外 2 6 が、縦軸に目的変数 y がプロットされている。

【0 0 5 6】

【外 2 4】

$\hat{\eta}$

【外 2 5】

$\hat{\eta}'z$

【外 2 6】

$\hat{\eta}'z$

図 7 は、相関行列で補正した E D R 方向外 2 7 へのスコア外 2 8 に対する目的変数 y の散布図である。図 4 と図 6，図 7 を比較してわかるように、本発明を用いて真の E D R 方向を推定することができる。尚、横軸に外 2 9 が、縦軸に目的変数 y がプロットされている。

【0 0 5 7】

【外 2 7】

$\hat{\eta}_N$

【外 2 8】

$\hat{\eta}'_N z$

【外 2 9】

$\hat{\eta}'_N z$

以下の表 1 は、真の E D R 方向へのスコアと推定された E D R 方向へのスコア

の相関係数の平均値および標準偏差（ $N = 50, 100, 500$ 、 $\rho = 0.0, 0.8$ として、100, 000回試行）、および推定されたEDR方向へのスコアと2値化された目的変数の相関係数の平均値および標準偏差（ $N = 50, 100, 500$ 、 $\rho = 0.0, 0.8$ として、100, 000回試行）を示す表である。ここで、 δ を2値化された目的変数とし、以下の数20に示された数式で表す。

【0058】

【表1】

N	$\rho = 0.0$		$\rho = 0.8$	
	$\text{Cor}(\hat{\eta}'z, \eta_0'z)$	$\text{Cor}(\hat{\eta}'z, \delta)$	$\text{Cor}(\hat{\eta}'z, \eta_0'z)$	$\text{Cor}(\hat{\eta}'z, \delta)$
50	0.936	0.803	0.921	0.769
	(0.039)	(0.034)	(0.032)	(0.039)
100	0.967	0.799	0.935	0.762
	(0.021)	(0.023)	(0.020)	(0.027)
500	0.993	0.798	0.946	0.758
	(0.004)	(0.010)	(0.007)	(0.012)

【数20】

$$\delta = \begin{cases} 1, & y \geq t, \\ -1, & y < t. \end{cases}$$

ここで、閾値 t は目的変数 y の中央値である。 $N = 50, 100, 500$ 、 $\rho = 0.0, 0.8$ と変化させて、それぞれ100, 000回解析したときの、相関係数の平均値、標準偏差を示している。上記表1より、真のEDR方向へのスコアと推定されたEDR方向へのスコアの相関係数は1に近く、分散が小さい値を示している。これにより、本発明を用いて真のEDR方向を推定できることがわかる。

【0059】

また、推定されたEDR方向へのスコアと2値化された目的変数の相関係数はサンプル数が大きくなってもあまり変化しないことを示している。これにより、データ数にあまり影響されずにEDR方向を推定することができることがわかる。

【0060】

【発明の効果】

本発明の第1の発明の効果は、単一指標モデルにおいて、データを基準化するとき分散共分散行列の逆行列を用いないことにある。この結果、大量変数についても、データを基準化することができる。その理由は、データの平均と分散のみを用いてデータを基準化するためである。

本発明の第2の発明の効果は、単一指標モデルにおいて、スライス数が2つのときのEDR方向を求めるのに、主成分分析をしないでEDR方向を求めることができることにある。この結果、大量変数からなる単一指標モデルにおいて、スライス数が2つのときのEDR方向を求めることが可能となり、計算速度も改善される。その理由は、平均ベクトルの差を計算するだけでEDR方向を求めることができるからである。

【0061】

以上の理由により、本手法は遺伝子発現解析用DNAチップやマイクロアレイなどの大量変数のデータに適用することができる。マイクロアレイのデータに適用するときは、目的変数 y は副作用などの表現形とし、 x はマイクロアレイにより得られる各遺伝子の発現量とする。このとき、得られたEDR方向の係数に関して、係数が大きい遺伝子Aと係数が小さい遺伝子Bでは、表現型に対して遺伝子Aの方が遺伝子Bよりも影響が大きい、つまり重要であることを示す。よって、係数の大きさにしたがって、表現型に対して重要な遺伝子を探索することが可能となる。

【図面の簡単な説明】

【図1】

本発明の第1の実施の形態の構成を示すブロック図である。

【図2】

本発明の第 1 の実施の形態の動作を示す流れ図である。

【図 3】

本発明の第 5 の実施の形態の構成を示すブロック図である。

【図 4】

モデルにより生成されたデータを表す散布図である。

【図 5】

$z(1)$ と $z(2)$ の散布図である。

【図 6】

推定された EDR 方向に対する目的変数の散布図である。

【図 7】

相関行列で補正した EDR 方向に対する目的変数の散布図である。

【図 8】

従来の技術の構成を示すブロック図である。

【図 9】

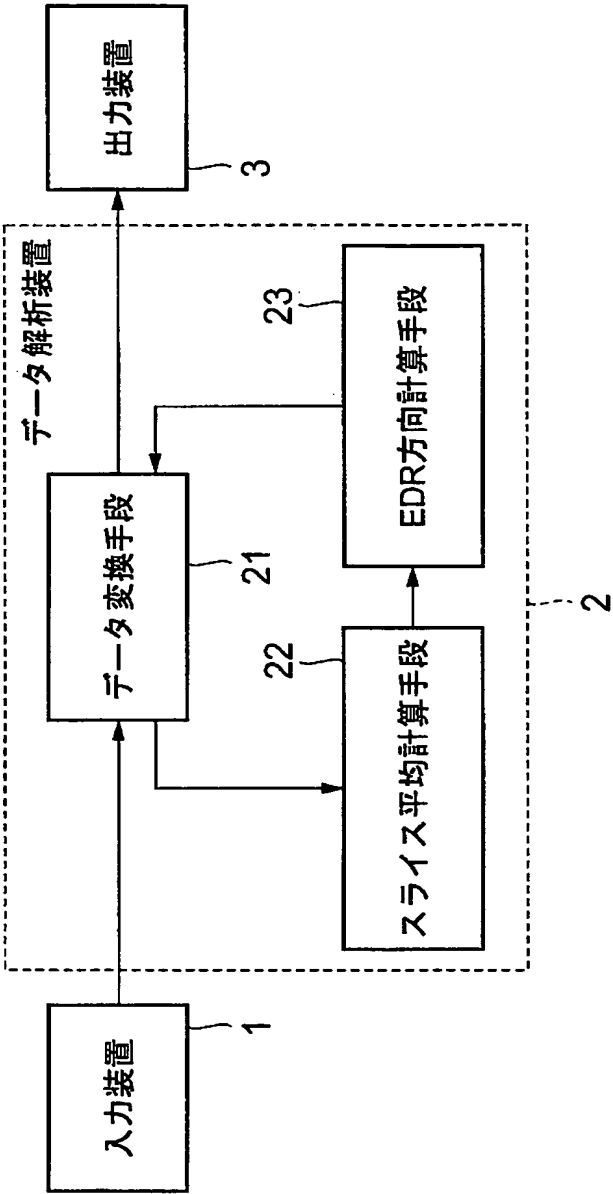
従来の技術の動作を示す流れ図である。

【符号の説明】

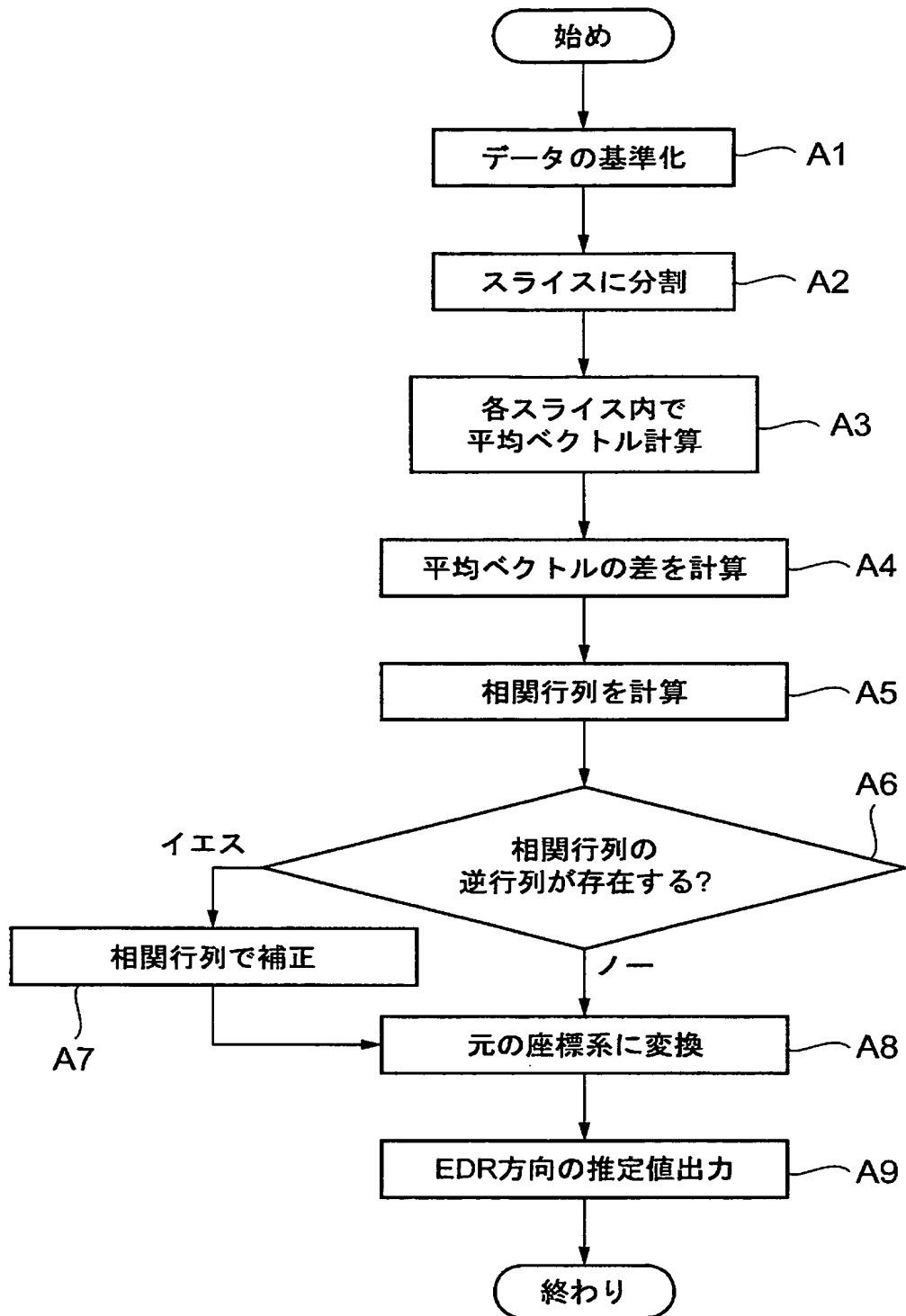
- 1 入力装置
- 2 データ解析装置
- 3 出力装置
- 4 記録媒体
- 5 データ解析装置
- 21 データ変換手段
- 22 スライス平均計算手段
- 23 EDR 方向計算手段
- 24 データ基準化手段
- 25 主成分分析手段

【書類名】 図面

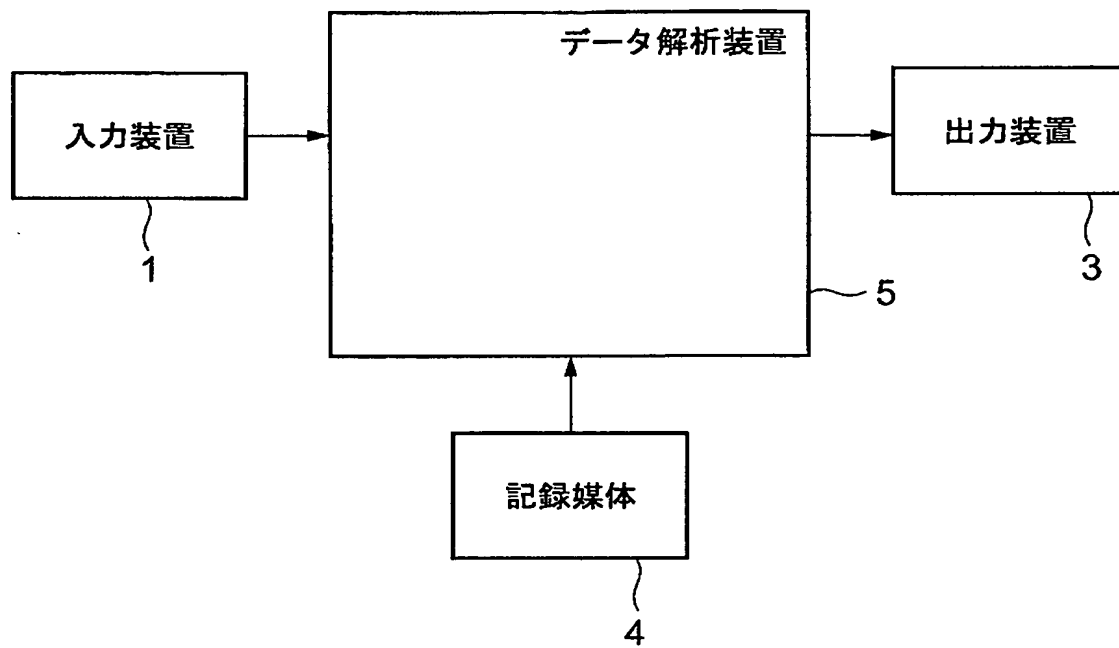
【図 1】



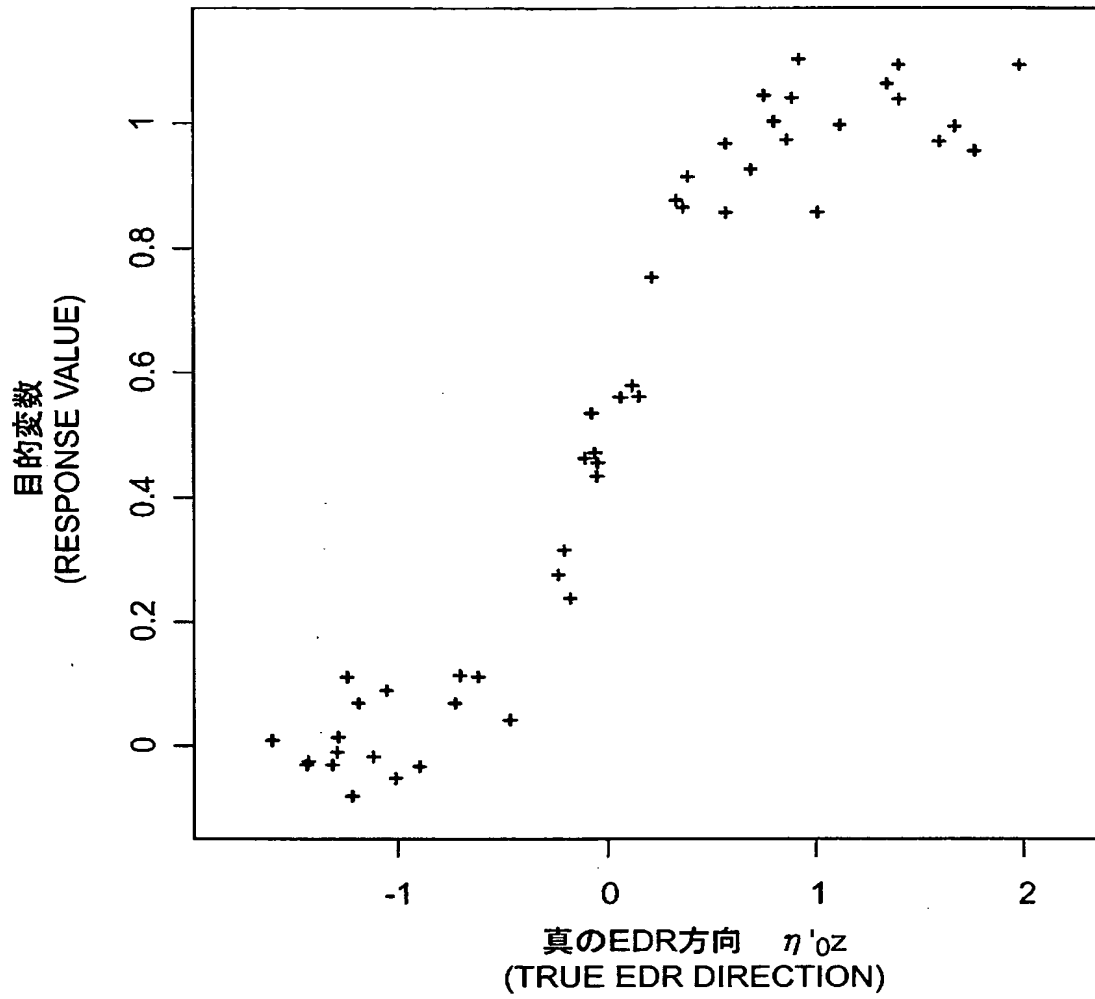
【図 2】



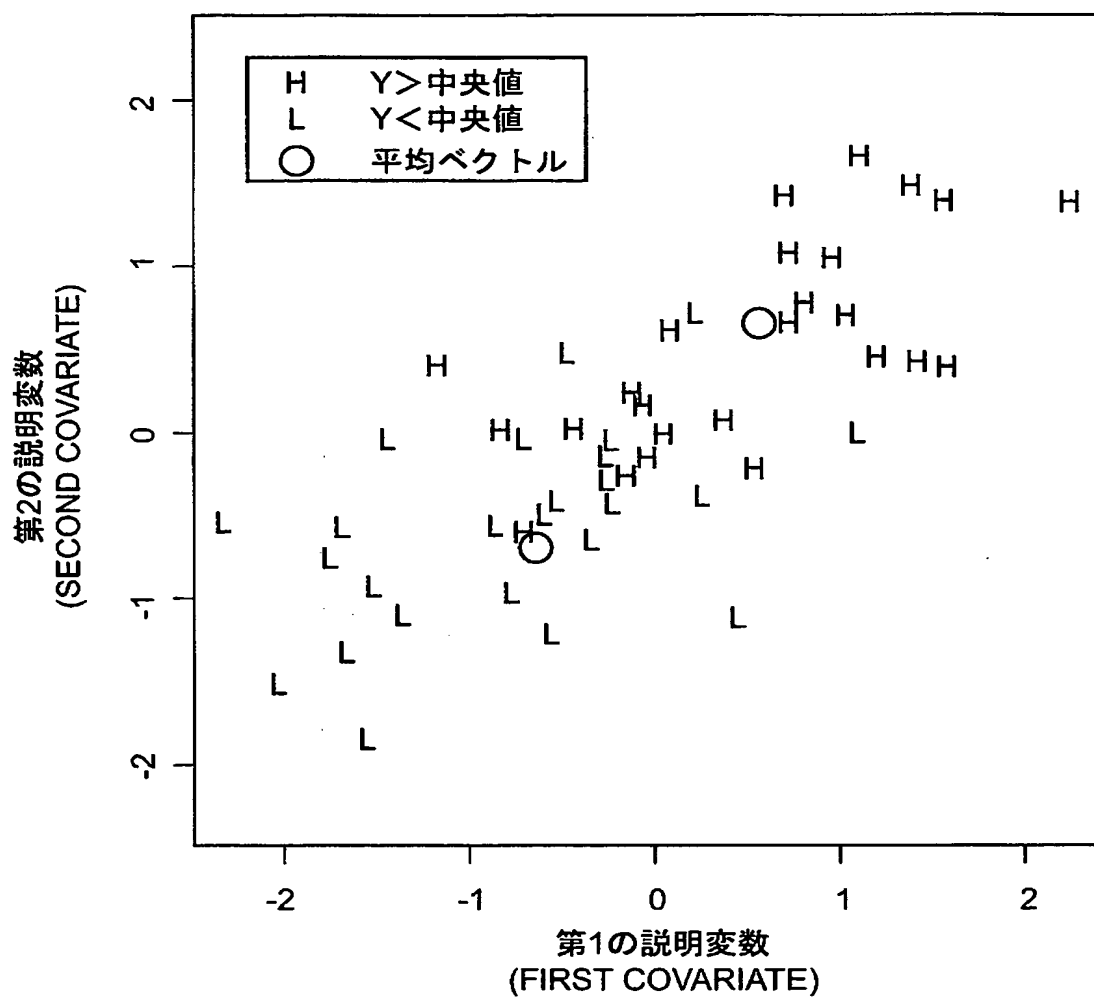
【図 3】



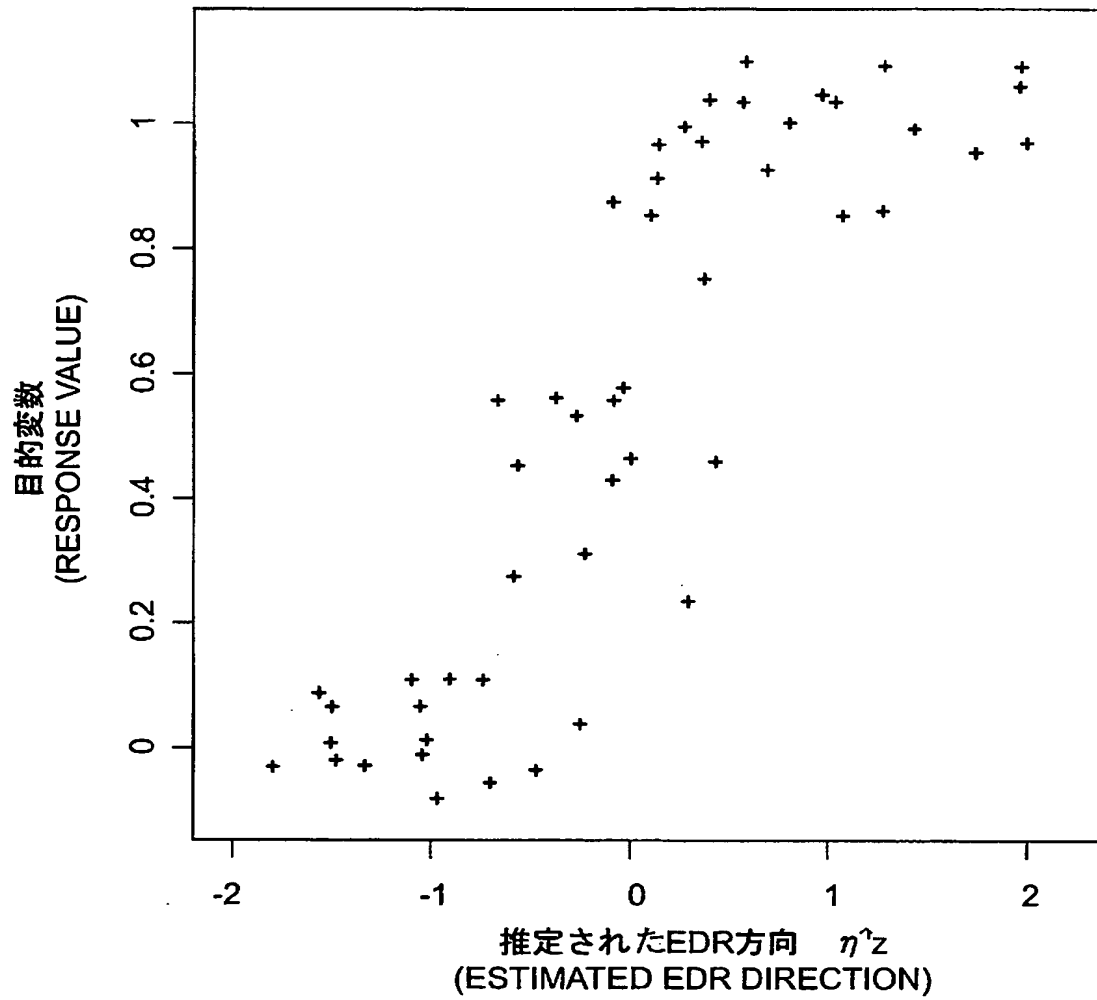
【図 4】



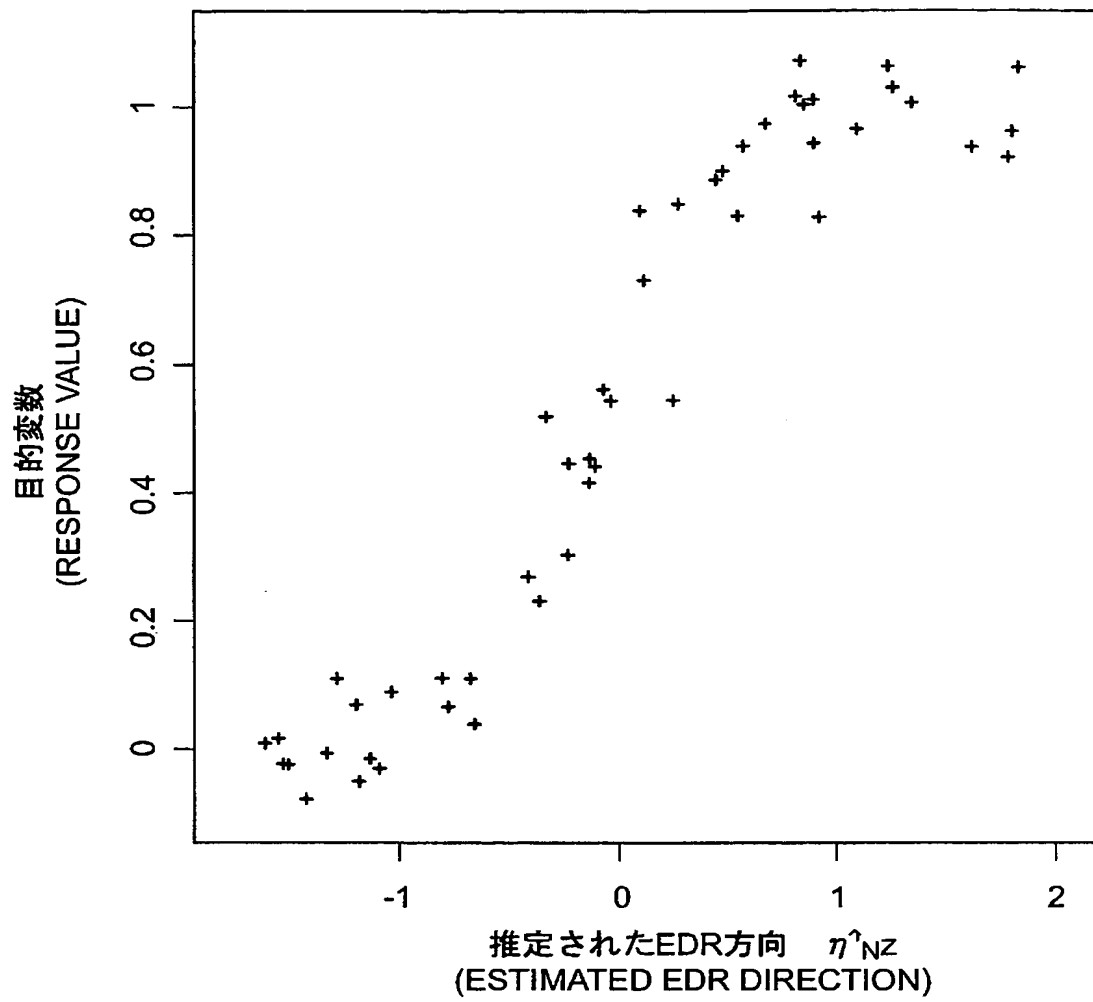
【図 5】



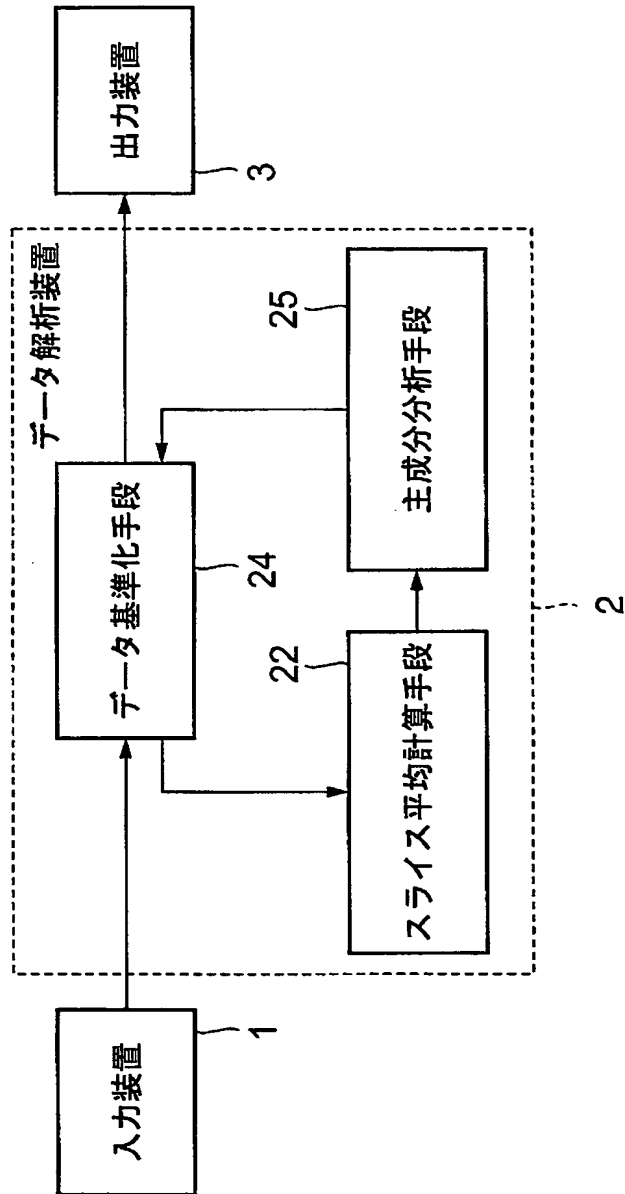
【図 6】



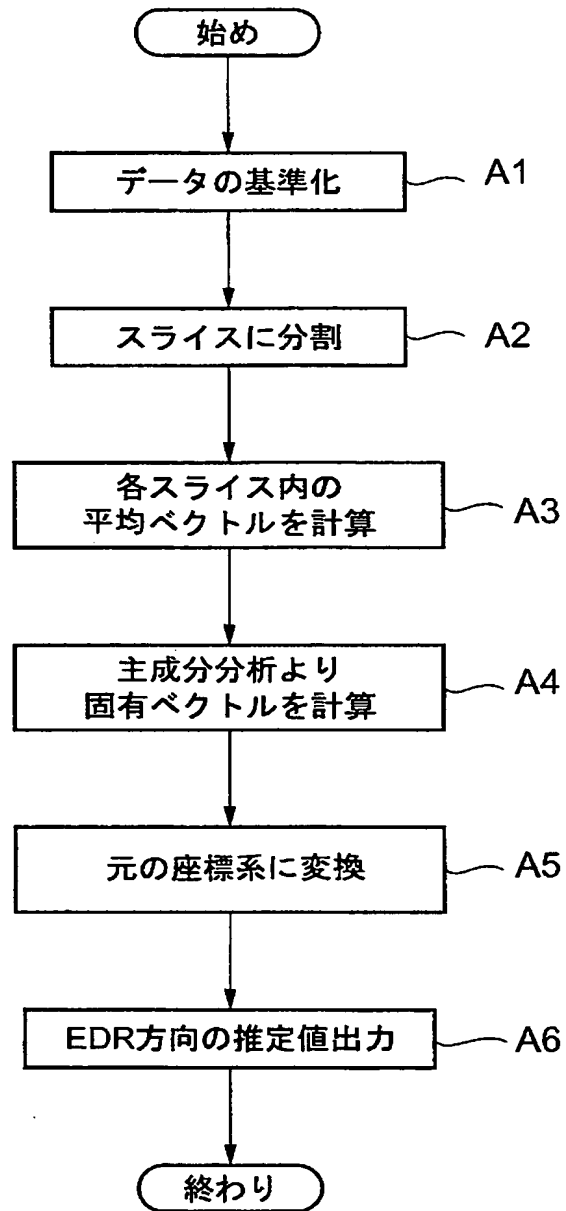
【図 7】



【図 8】



【図 9】



【書類名】 要約書

【要約】

【課題】 大量変数からなる単一指標モデルにおいて、分散共分散行列の逆行列および主成分分析を用いることなく、単純な計算でEDR方向を推定する。

【解決手段】 データ変換手段21は入力装置1から、目的変数と説明変数の組からなる解析対象データを受け取り、説明変数を基準化してスライス平均計算手段22に送る。スライス平均計算手段22はデータを目的変数の中央値を基準として2つのスライスに分割して、スライスごとに説明変数の平均ベクトルを計算する。計算された平均ベクトルはEDR方向計算手段23に送られる。EDR方向計算手段23はスライスごとの平均ベクトルの差を計算し、EDR方向を推定する。また、説明変数の相関行列の逆行列を用いて、推定されたEDR方向を補正する。推定されたEDR方向および補正されたEDR方向はデータ変換手段21に送られ、データ変換手段21において元の座標系に変換する。

【選択図】 図1

特願 2 0 0 3 - 0 4 9 2 2 3

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 4 2 3 7]

1. 変更年月日

1 9 9 0 年 8 月 2 9 日

[変更理由]

新規登録

住 所

東京都港区芝五丁目 7 番 1 号

氏 名

日本電気株式会社

特願 2 0 0 3 - 0 4 9 2 2 3

出 願 人 履 歴 情 報

識別番号

[5 0 3 0 7 7 1 6 5]

1. 変更年月日

2 0 0 3 年 2 月 2 6 日

[変更理由]

新規登録

住 所

広島県廿日市市宮園 9 丁目 1 の 7

氏 名

大瀧 慈